Chapter VI

Mining Geo-Referenced Databases: A Way to Improve Decision-Making

Maribel Yasmina Santos, University of Minho, Portugal

Luís Alfredo Amaral, University of Minho, Portugal

Abstract

Knowledge discovery in databases is a process that aims at the discovery of associations within data sets. The analysis of geo-referenced data demands a particular approach in this process. This chapter presents a new approach to the process of knowledge discovery, in which qualitative geographic identifiers give the positional aspects of geographic data. Those identifiers are manipulated using qualitative reasoning principles, which allows for the inference of new spatial relations required for the data mining step of the knowledge discovery process. The efficacy and usefulness of the implemented system — PADRÃO — has been tested with a bank dataset. The results obtained support that traditional knowledge discovery systems, developed for relational databases and not having semantic knowledge linked to spatial data, can be used in the process of knowledge discovery in geo-referenced databases, since some of this semantic knowledge and the principles of qualitative spatial reasoning are available as spatial domain knowledge.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Introduction

Knowledge discovery in databases is a process that aims at the discovery of associations within data sets. Data mining is the central step of this process. It corresponds to the application of algorithms for identifying patterns within data. Other steps are related to incorporating prior domain knowledge and interpretation of results.

The analysis of geo-referenced databases constitutes a special case that demands a particular approach within the knowledge discovery process. Geo-referenced data sets include allusion to geographical objects, locations or administrative sub-divisions of a region. The geographical location and extension of these objects define implicit relationships of spatial neighborhood. The data mining algorithms have to take this spatial neighborhood into account when looking for associations among data. They must evaluate if the geographic component has any influence in the patterns that can be identified.

Data mining algorithms available in traditional knowledge discovery tools, which have been developed for the analysis of relational databases, are not prepared for the analysis of this spatial component. This situation led to: (i) the development of new algorithms capable of dealing with spatial relationships; (ii) the adaptation of existing algorithms in order to enable them to deal with those spatial relationships; (iii) the integration of the capabilities for spatial analysis of spatial database management systems or geographical information systems with the tools normally used in the knowledge discovery process.

Most of the geographical attributes normally found in organizational databases (e.g., addresses) correspond to a type of spatial information, namely qualitative, which can be described using indirect positioning systems. In systems of spatial referencing using geographic identifiers, a position is referenced with respect to a real world location defined by a real world object. This object is termed a *location*, and its identifier is termed a *geographic identifier*. These geographic identifiers are very common in organizational databases, and they allow the integration of the spatial component associated with them in the process of knowledge discovery.

This chapter presents a new approach to the analysis of geo-referenced data. It is based on qualitative spatial reasoning strategies, which enable the integration of the spatial component in the knowledge discovery process. This approach, implemented in the PADRÃO system, allowed the analysis of geo-referenced databases and the identification of implicit relationships existing between the geo-spatial and non-spatial data.

The following sections, in outline, include: (i) an overview of the process of knowledge discovery and its several phases. The approaches usually followed in the analysis of geo-referenced databases are also presented; (ii) a description of qualitative spatial reasoning presenting its principles and the several spatial relations — direction, distance and topology. For the relations, an integrated spatial reasoning system was constructed and made available in the Spatial Knowledge Base of the PADRÃO system. The rules stored enable the inference of new spatial relations needed in the data mining step of the knowledge discovery process; (iii) a presentation of the PADRÃO system describing its architecture and its implementation achieved through the adoption of several technologies. This section continues with the analysis of a geo-referenced database, based on

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

the several steps of the knowledge discovery process considered by the PADRÃO system; and (iv) a conclusion with some comments about the proposed research and its main advantages.

Knowledge Discovery in Databases

Large amounts of operational data concerning several years of operation are available, mainly from middle-large sized organizations. Knowledge discovery in databases is the key to gaining access to the strategic value of the organizational knowledge stored in databases; for use in daily operations, general management and strategic planning.

Knowledge Discovery Process

Knowledge Discovery in Databases (KDD) is a complex process concerning the discovery of relationships and other descriptions from data. Data mining refers to the application algorithms used to extract patterns from data without the additional steps of the KDD process, e.g., the incorporation of appropriate prior knowledge and the interpretation of results (Fayyad & Uthurusamy, 1996).

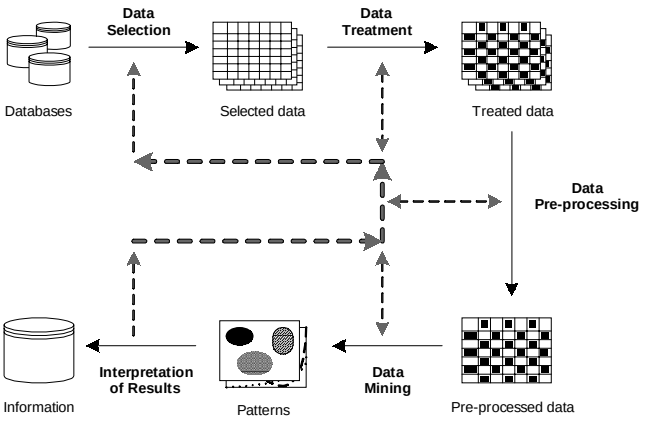
Different tasks can be performed in the knowledge discovery process and several techniques can be applied for the execution of a specific task. Among the available tasks are *classification*, *clustering*, *association*, *estimation* and *summarization*. KDD applications integrate a variety of data mining algorithms. The performance of each technique (algorithm) depends upon the task to be carried out, the quality of the available data and the objective of the discovery. The most popular Data Mining algorithms include *neural networks*, *decision trees*, *association rules* and *genetic algorithms* (Han & Kamber, 2001).

The steps of the KDD process (*Figure 1*) include data selection, data treatment, data preprocessing, data mining and interpretation of results. This process is interactive, because it requires user participation, and iterative, because it allows for going back to a previous phase and then proceeding forward with the knowledge discovery process. The steps of the KDD process are briefly described:

- *Data Selection.* This step allows for the selection of relevant data needed for the execution of a defined data mining task. In this phase the minimum sub-set of data to be selected, the size of the sample needed and the period of time to be considered must be evaluated.
- *Data Treatment.* This phase concerns with the cleaning up of selected data, which allows for the treatment of corrupted data and the definition of strategies for dealing with missing data fields.
- *Data Pre-Processing.* This step makes possible the reduction of the sample destined for analysis. Two tasks can be carried out here: (i) the reduction of the

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 1. Knowledge Discovery Process



number of rows or, (ii) the reduction of the number of columns. In the reduction of the number of rows, data can be generalized according to the defined hierarchies or attributes with continuous values can be transformed into discreet values according to the defined classes. The reduction of the number of columns attempts to verify if any of the selected attributes can now be omitted.

- *Data Mining.* Several algorithms can be used for the execution of a given data mining task. In this step, various available algorithms are evaluated in order to identify the most appropriate for the execution of the defined task. The selected one is applied to the relevant data in order to find implicit relationships or other interesting patterns that exist in the data.
- *Interpretation of Results.* The interpretation of the discovered patterns aims at evaluating their utility and importance with respect to the application domain. It may be determined that relevant attributes were ignored in the analysis, thus suggesting that the process should be repeated.

Knowledge Discovery in Spatial Databases

The main recognized advances in the area of KDD (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996) are related with the exploration of relational databases. However, in most organizational databases there exists one dimension of data, the *geographic* (associated with addresses or post-codes), the semantic of which is not used by traditional KDD systems.

Knowledge Discovery in Spatial Databases (KDSD) is related with "*the extraction of interesting spatial patterns and features, general relationships that exist between*

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

spatial and non-spatial data, and other data characteristics not explicitly stored in spatial databases" (Koperski & Han, 1995).

Spatial database systems are relational databases with a concept of spatial location and spatial extension (Ester, Kriegel & Sander, 1997). The explicit location and extension of objects define implicit relationships of spatial neighborhood. The major difference between knowledge discovery in relational databases and KDSD is that the neighbor attributes of an object may influence the object itself and, therefore, must be considered in the knowledge discovery process. For example, a new industrial plant may pollute its neighborhood entities depending on the distance between the objects (regions) and the major direction of the wind. Traditionally, knowledge discovery in relational databases does not take into account this spatial reasoning, which motivates the development of new algorithms adapted to the spatial component of spatial data.

The main approaches in KDSD are characterized by the development of new algorithms that treat the position and extension of objects mainly through the manipulation of their coordinates. These algorithms are then implemented, thus extending traditional KDD systems in order to accommodate them. In all, a quantitative approach is used in the spatial reasoning process although the results are presented using qualitative identifiers.

Lu, Han & Ooi (1993) proposed an attribute-oriented induction approach that is applied to spatial and non-spatial attributes using conceptual hierarchies. This allows the discovery of relationships that exist between spatial and non-spatial data. A spatial concept hierarchy represents a successive merge of neighborhood regions into large regions. Two learning algorithms were introduced: (i) non-spatial attribute-oriented induction, which performs generalization on non-spatial data first, and (ii) spatial hierarchy induction, which performs generalization on spatial data first. In both approaches, the classification of the corresponding spatial and non-spatial data is performed based on the classes obtained by the generalization. Another peculiarity of this approach is that the user must provide the system with the relevant data set, the concept hierarchies, the desired rule form and the learning request (specified in a syntax similar to SQL – Structured Query Language).

Koperski & Han (1995) investigated the utilization of interactive data mining for the extraction of spatial association rules. In their approach the spatial and non-spatial attributes are held in different databases, but once the user identifies the attributes or relationships of interest, a selection process takes place and a unified database is created. An algorithm, implemented for the discovery of spatial association rules, analyzes the stored data. The rules obtained represent relationships between objects, described using spatial predicates like *adjacent to* or *close to*.

These approaches are two examples of the efforts made in the area of KDSD. One approach uses two different databases, storing spatial and non-spatial data separately. Once the user identifies the attributes of interest, an interface between the two databases ensures the selection and treatment of data without the creation of a new integrated repository. The other approach also requires two different databases, but the selection phase leads to the creation of a unified database where the analysis of data takes place. In both approaches new algorithms were implemented and the user is asked for the specification of the relevant attributes and the type of results expected.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Two approaches for the analysis of spatial data with the aim of knowledge discovery have been presented. Independently of the adopted approach, several tasks can be performed in this process, among them: *spatial characterization*, *spatial classification*, *spatial association* and *spatial trends analysis* (Koperski & Han, 1995; Ester, Frommelt, Kriegel & Sander, 1998; Han & Kamber, 2001).

A *spatial characterization* corresponds to a description of the spatial and non-spatial properties of a selected set of objects. This task is achieved analyzing not only the properties of the target objects, but also the properties of their neighbors. In a characterization, the relative frequency of incidence of a property in the selected objects, and their neighbors, is different from the relative frequency of the same property verified in the remaining of the database (Ester, Frommelt, Kriegel & Sander, 1998). For example, the incidence of a particular disease can be higher in a set of regions closest or holding a specific industrial complex, showing that a possible *cause-effect* relationship exists between the disease and the industry pollution.

Spatial classification aims to classify spatial objects based on the spatial and non-spatial features of these objects in a database. The result of the classification, a set of rules that divides the data into several classes, can be used to get a better understanding of the relationships among the objects in the database and to predict characteristics of new objects (Han, Tung & He, 2001; Han & Kamber, 2001). For example, regions can be classified into *rich* or *poor* according to the average family income or any other relevant attribute present in the database.

Spatial association permits the identification of spatial-related association rules from a set of data. An association rule shows the frequently occurring patterns of a set of data items in a database. A spatial association rule is a rule of the form "$X \rightarrow Y$ (s%, c%)," where X and Y are sets of spatial and non-spatial predicates (Koperski & Han, 1995). In an association rule, s represents the support of the rule, the probability that X and Y exist together in the data items analyzed, while c indicates the confidence of the rule, i.e., the probability that Y is true under the condition of X. For example, the spatial association rule "$\text{is\_a}(x, \text{House}) \wedge \text{close\_to}(x, \text{Beach}) \rightarrow \text{is\_expensive}(x)$" states that houses which are close to the beach are expensive.

A *spatial trend* (Ester, Frommelt, Kriegel & Sander, 1998) describes a regular change of one or more non-spatial attributes when moving away from a particular spatial object. Spatial trend analysis allows for the detection of changes and trends along a spatial dimension. Examples of spatial trends are the changes in the economic situation of a population when moving away from the center of a city or the trend of change of the climate with the increasing distance from the ocean (Han & Kamber, 2001).

After the presentation of two approaches and some of the most popular tasks associated with the analysis of spatial data with the aim of knowledge discovery, this chapter posits a new approach to the process of KDSD (more specifically in geo-referenced datasets). This approach integrates qualitative principles in the spatial reasoning system used in the knowledge discovery process. Since the use of coordinates for the identification of a spatial object is not always needed, this work investigates how traditional KDD systems (and their generic data mining algorithms) can be used in KDSD.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Qualitative Spatial Reasoning

Human beings use qualitative identifiers extensively to simplify reality and to perform spatial reasoning more efficiently. *Spatial reasoning* is the process by which information about objects in space and their relationships are gathered through measurement, observation or inference and used to arrive at valid conclusions regarding the relationships of the objects (Sharma, 1996). *Qualitative spatial reasoning* (Abdelmoty & El-Geresy, 1995) is based on the manipulation of qualitative spatial relations, for which composition[1] tables facilitate reasoning, thereby allowing the inference of new spatial knowledge.

Spatial relations have been classified into several types (Frank, 1996; Papadias & Sellis, 1994), including *direction relations* (Freksa, 1992) (that describe order in space), *distance relations* (Hernández, Clementini & Felice, 1995) (that describe proximity in space) and *topological relations* (Egenhofer, 1994) (that describe neighborhood and incidence). Qualitative spatial relations are specified by using a small set of symbols, like *North, close, etc.*, and are manipulated through a set of inference rules.

The inference of new spatial relations can be achieved using the defined qualitative rules, which are compiled into a composition table. These rules allow for the manipulation of the qualitative identifiers adopted. For example, knowing the facts, A North, very far from B and B Northeast, very close to C, it is possible, by consulting the composition table for integrated direction and distance spatial reasoning (presented later), to infer the relationship that exists between A and C, that is A North, very far from C.

The inference rules can be constructed using quantitative methods (Hong, 1994) or by manipulating qualitatively the set of identifiers adopted (Frank, 1992; Frank, 1996), an approach that requires the definition of axioms and properties for the spatial domain.

Later in this section the construction of the qualitative spatial reasoning system used by PADRÃO is presented. The qualitative system integrates *direction*, *distance* and *topological* spatial relations. Its conception was achieved based on the work developed by Hong (1994) and Sharma (1996). The application domain in which this qualitative reasoning system will be used is characterized by objects that represent administrative subdivisions.

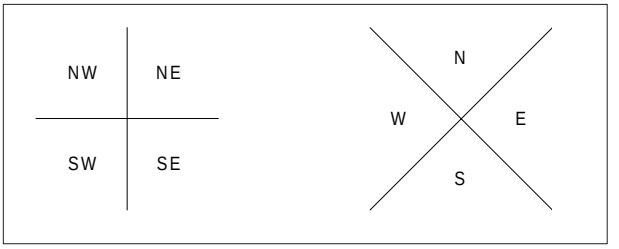
Direction Spatial Relations

Direction relations describe where objects are placed relative to each other. Three elements are needed to establish an orientation: two objects and a fixed point of reference (usually the North Pole) (Frank, 1996; Freksa, 1992). Cardinal directions can be expressed using numerical values specifying degrees (0º, 45º...) or using qualitative values or symbols, such as North or South, which have an associated acceptance region. The regions of acceptance for qualitative directions can be obtained by projections (also known as half-planes) or by cone-shaped regions (*Figure 2*).

A characteristic of the cone-shaped system is that the region of acceptance increases with distance, which makes it suitable for the definition of direction relations between

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 2. Direction Relations Definition by Projection and Cone-Shaped Systems



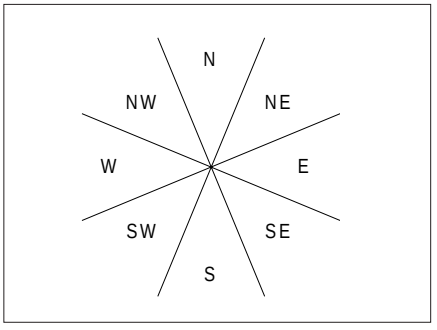
extended objects[2] (Sharma, 1996). It also allows for the definition of finer resolutions, thus permitting the use of eight (*Figure 3*) or 16 different qualitative directions. This model uses triangular acceptance areas that are drawn from the *centroid* of the reference object towards the primary object (in the spatial relation A North B, B represents the reference object, while A constitutes the primary object).

Distance Spatial Relations

Distances are quantitative values determined through measurements or calculated from known coordinates of two objects in some reference system. The frequently used definition of distance can be achieved using the Euclidean geometry and Cartesian coordinates. In a two-dimensional Cartesian system, it corresponds to the length of the shortest possible path (a straight line) between two objects, which is also known as the Euclidean distance (Hong, 1994). Usually a metric quantity is mapped onto some qualitative indicator such as very close or far for human common-sense reasoning (Hernández et al., 1995).

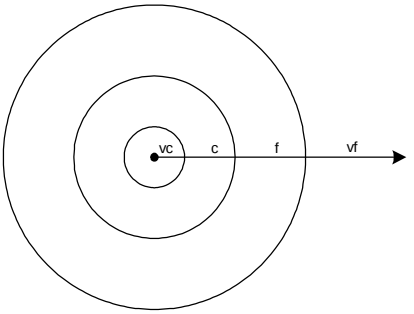
Qualitative distances must correspond to a range of quantitative values specified by an interval and they should be ordered so that comparisons are possible. The adoption of

Figure 3. Cone-Shaped System with Eight Regions of Acceptance



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 4. Qualitative Distances Intervals

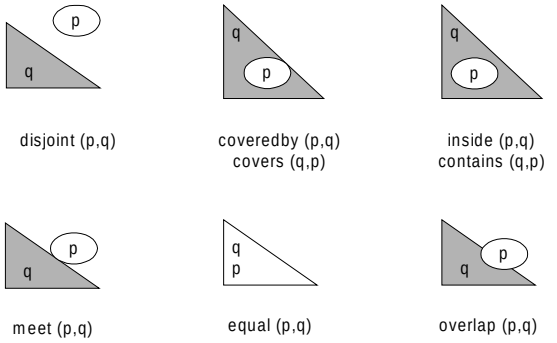


the qualitative distances very close – vc, close – c, far – f and very far – vf, intuitively describe distances from the nearest to the furthest. An order relationship exists among these relations, where a lower order (vc) relates to shorter quantitative distances and a higher order (vf) relates to longer quantitative distances (Hong, 1994). The length of each successive qualitative distance, in terms of quantitative values, should be greater or equal to the length of the previous one (*Figure 4*).

Topological Spatial Relations

Topological relations are those relationships that are invariant under continuous transformations of space such as rotation or scaling. There are eight topological relations that can exist between two planar regions without holes[3]: disjoint, contains, inside, equal, meet, covers, covered by and overlap (*Figure 5*). These relations can be defined considering intersections between the two regions, their boundaries and their complements (Egenhofer, 1994). These eight relations, which can exist between two spatial regions without holes, will be the exclusive focus of topological relations in this chapter.

Figure 5. Topological Spatial Relation



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

In some exceptional cases, the geographic space cannot be characterized, in topological terms, with reference to the eight topological primitives presented above. One of these cases is related with application domains in which the geographic regions addressed are administrative subdivisions. Administrative subdivisions, represented in this work by full planar graphs[4], can only be related through the topological primitives disjoint, meet and contains (and the corresponding inverse inside), since they cannot have any kind of overlapping. The topological primitives used in this chapter are disjoint and meet, since the implemented qualitative inference process only considers regions at the same geographic hierarchical level.

Integrated Spatial Reasoning

Integrated reasoning about qualitative directions necessarily involves qualitative distances and directions. Particularly in objects with extension, the size and shape of objects and the distance between them influence the directions. One of the ways to determine the direction and distance[5] between regions is to calculate them from the *centroids* of the regions. The extension of the geographic entities is somehow implicit in the topological primitive used to characterize their relationships.

Integration of Direction and Distance

An example of *integrated spatial reasonin*g about qualitative distances and directions is as follows. The facts A is very far from B and B is very far from C do not facilitate the inference of the relationship that exists between A and C. A can be very close or close to C, or A may be far or very far from C, depending on the orientation between B and C.

For the integration of qualitative distances and directions the adoption of a set of identifiers is required, which allows for the identification of the considered directions and distances and their respective intervals of validity. Hong (1994) analyzed some possible combinations for the number of identifiers and the geometric patterns that should characterize the distance intervals. The *localization system* (*Figure 6*) suggested by Hong is based on eight symbols for direction relations (North, Northeast, East, Southeast, South, Southwest, West, Northwest) and four symbols for the identification of the distance relations (very close, close, far and very far).

In the case of direction relations, for the cone-shaped system with eight acceptance regions, the quantitative intervals adopted were: [337.5, 22.5), [22.5, 67.5), [67.5, 112.5), [112.5, 157.5), [157.5, 202.5), [202.5, 247.5), [247.5, 292.5), [292.5, 337.5) from North to Northwest respectively.

The definition of the validity interval for each distance identifier must obey some rules (Hong, 1994). In these systems, as can be seen in *Table 1*, there should exist a constant ratio (ratio = length ($dist_i$)/length ($dist_{i-1}$)) relationship between the lengths of two neighboring intervals. The presented simulated intervals allow for the definition of new distance intervals by magnification of the original intervals. For example, the set of values for ratio 4[6] can be increased by a factor of 10 supplying the values $dist_0$ (0, 10], $dist_1$

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 6. Integration of Direction and Distance Spatial Relations

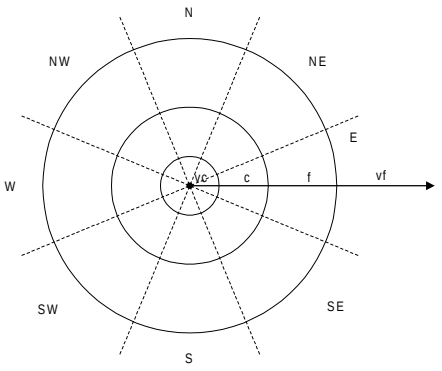


Table 1. Simulated Intervals for Four Symbolic Distance Values

Ratio	$dist_0$	$dist_1$	$dist_2$	$dist_3$
1	(0, 1]	(1, 2]	(2, 3]	(3, 4]
2	(0, 1]	(1, 3]	(3, 7]	(7, 15]
3	(0, 1]	(1, 4]	(4, 13]	(13, 40]
4	(0, 1]	(1, 5]	(5, 21]	(21, 85]
5	(0, 1]	(1, 6]	(6, 31]	(31, 156]
…	…	…	…	…

(10, 50], $dist_2$ (50, 210] and $dist_3$ (210, 850]. Since the same scale magnifies all intervals and quantitative distance relations, the qualitative compositions will remain the same, regardless of the scaled value.

It is important to know that the number of distance symbols used and the ratio between the quantitative values addressed by each interval play an important role in the robustness of the final system, i.e., in the validity of the composition table for the inference of new spatial relations (Hong, 1994).

The final composition table, a 32x32 matrix for the localization system adopted, was constructed following the suggestions made by Hong (1994) and it is presented in this work through an iconic representation (*Figure 7*). This matrix represents part of the knowledge needed for the inference of new spatial information in the localization system used. Due to its great size, *Figure 8* exhibits an extract of the final matrix. An example of the composition operation: suppose that A North, close B and that B Southeast, very close C. Consulting the composition table (this example is marked in *Figure 8* with two traced arrows) it is possible to identify the relation that exists between A and C: A North, close C. For the particular case of the composition of opposite directions with equal qualitative distances, the system is unable to identify the direction between the objects. For this

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 7. Graphical Representation of Direction and Distance Integration

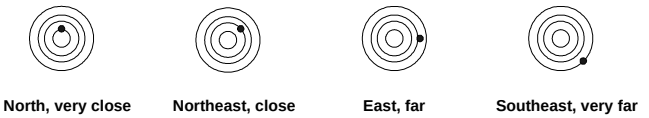
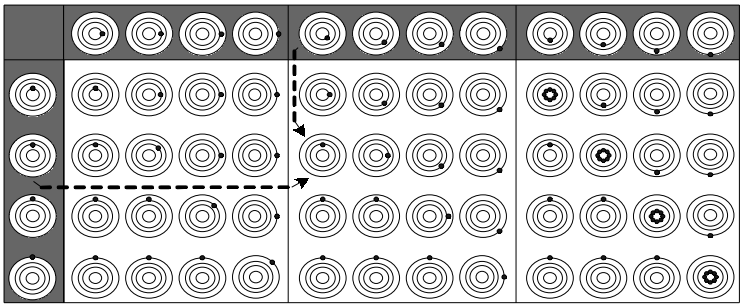


Figure 8. Extract of the Final Composition Table — Integration of Direction and Distance



reason, the composition of these particular cases presents all the qualitative directions as possible results of the inference (*Figure 8*).

Integration of Direction and Topology

The relative position of two objects in the bi-dimensional space can be achieved through the dimension and orientation of the objects. Looking at each of these characteristics separately implies two classes of spatial relations: *topological*, which ignores orientations in space; and *direction* that ignores the extension of the objects.

The integration of these two kinds of spatial relations enables the definition of a system for qualitative spatial reasoning that describes the relative position existing between the objects and how the limits (frontiers) of them are related.

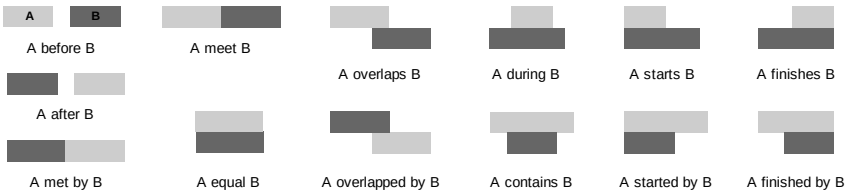
Sharma (1996) integrated direction and topological spatial relations using the principles of qualitative temporal reasoning defined by Allen (1983). The approach undertaken by Sharma (1996) was possible through the adaptation of the temporal principles to the spatial domain. The 13 temporal primitives (Allen, 1983) are: before, after, during, contain, overlap, overlapped by, meet, met by, start, started by, finish, finished by and equal (*Figure 9*).

The temporal primitives (that are one-dimensional) were analyzed by Sharma (1996) along two dimensions (axes *xx* and *yy*) allowing their use in the spatial domain (restricted in this case to a two-dimensional space).

The construction of the composition tables was facilitated by the knowledge representation framework adopted for the integration of direction and topology. Topological

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 9. Temporal Primitives



Adapted from Allen (1983, p. 835)

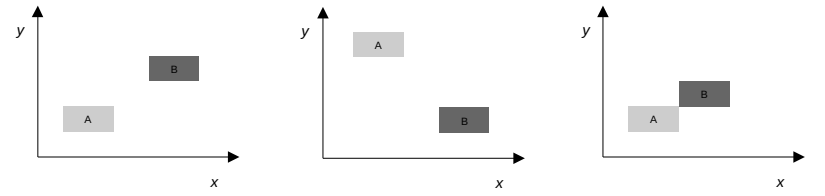
relations are independent of the order existing between the objects when analyzed along a given axis. Direction relations depend on the order and are defined by verifying the objects position along a specific axis.

The representation of each pair (direction, topology) is accomplished through temporal primitives. The transformation of the one-dimensional characteristics to the two-dimensional space is achieved analyzing the pair of temporal primitives that represent the behavior of the pair (direction, topology) along *x* and *y* (*Figure 10* supplies three examples of selection of the appropriate pair of temporal primitives, verifying the position of A and B along *x* and along *y*, for the characterization of the pair (direction, topology).

Restricting the integration domain to objects that represent administrative subdivisions without overlap between them, the two topological relations considered were disjoint and meet. These two topological relationships can be represented by the temporal primitives before and meet, and by the corresponding inverses (after and met by). Attending to the direction relations, all the temporal primitives defined by Allen (1983) can be used in their characterization. *Figure 11* shows how the temporal primitives are used in the definition of a particular direction relation.

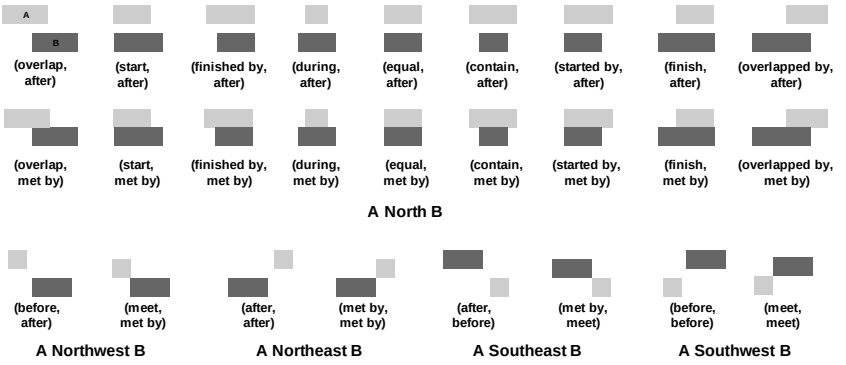
For the identification of the inference rules it is necessary to identify the temporal primitives that characterize each pair (direction, topology) and then do their composition

Figure 10. Integration of Direction and Topological Relations



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 11. Interval Relations for Direction Relations Representation



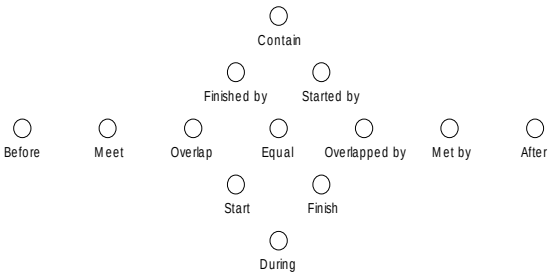
Adapted from Sharma (1996, p. 83)

to achieve the result. *Table 2* presents an extract of the composition table for the temporal domain. This table, graphically presented using the notation showed in *Figure 12* , will be afterwards used for the spatial domain.

The composition of pairs of relations (direction, topology) is performed consulting *Table 2*. An example of the composition[7] operation for the spatial domain is the composition of the pair (Northeast, disjoint) with the pair (Northeast, disjoint). The result of the composition is achieved by the steps:

(Northeast, disjoint) ; (Northeast, disjoint) = *(after, after) ; (after, after)*
= *(after; after) x (after; after)*
= *(after) x (after)*
= *(after, after)*
= *(Northeast, disjoint)*

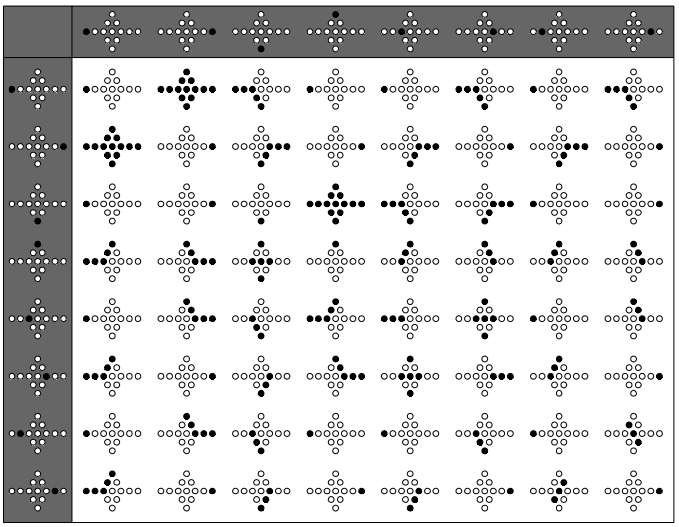
Figure 12. Temporal Relations — Graphical Representation



Notation suggested by Sharma (1996)

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Table 2. An Extract of the Composition Table for Temporal Intervals



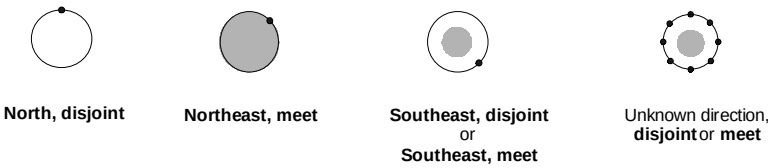
Adapted from Allen (1983, p. 836)

Following this composition process, Sharma obtained the several composition tables that integrate direction with the several topological pairs disjoint;disjoint, disjoint;meet, meet;disjoint and meet;meet. *Figure 13* presents the graphical symbols used in this chapter to represent the integration of direction and topology. *Table 3* shows one of the composition tables of Sharma, integrating direction with the topological pair disjoint;disjoint.

Integration of Direction, Distance and Topological Spatial Relations

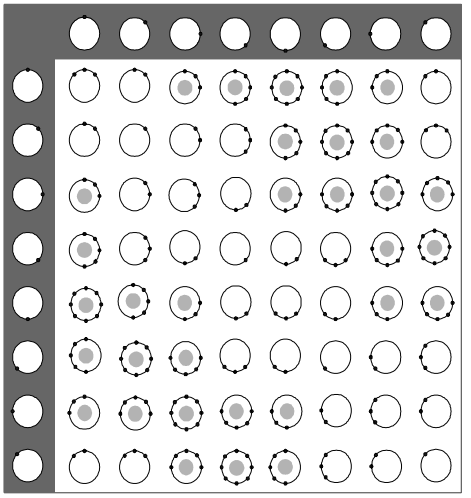
With the integration of direction and distance spatial relations a set of inference rules were obtained. These rules present a unique pair (direction, distance) as outcome, with the exception of the result of the composition of pairs with opposite directions and equal qualitative distances. In the integration of direction and topological spatial relations some improvements can be achieved, since several inference rules present as the result a set of outcomes.

Figure 13. Graphical Representation of Direction and Topological Spatial Relations



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Table 3. Composition Table for the Integration of Direction with the Topological Pair disjoint;disjoint



Adapted from, Sharma (1996, p. 117)

Looking at the work developed by Hong and Sharma it was realized that the integration of the three types of spatial relations, direction, distance and topology, would lead to more accurate composition tables.

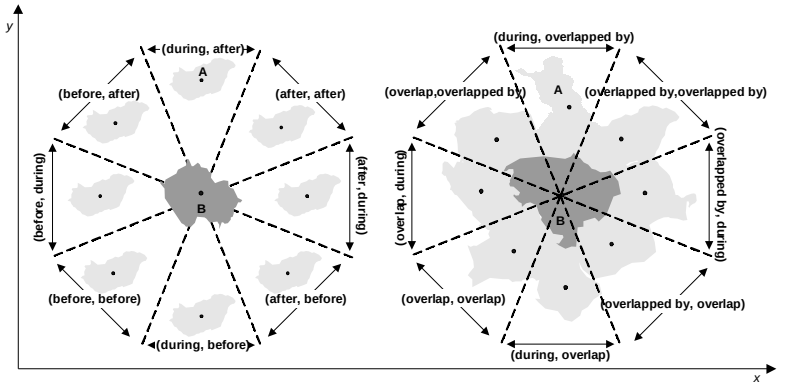
Since Hong adopted a cone-shaped system in the definition of the direction relations, and Sharma used a projection-based system for the same task, the integration of the three types of spatial relations was preceded by the adaptation[8] of the principles used by Sharma and the construction of new composition tables for the integration of direction and topology.

For the characterization of the integration of direction and topological relations, for the particular case of administrative subdivisions, new temporal pairs were defined, which allowed for the identification of new inference rules. *Figure 14* shows the several pairs of temporal primitives adopted according to the direction relations and the topological primitives disjoint and meet.

The adoption of the temporal intervals shown in *Figure 14* was motivated by the fact that administrative subdivisions have irregular limits, which impose several difficulties in the identification of the *correct* direction between two regions. Sometimes the *centroid* is positioned in a place that suggests one direction, although the administrative region may have parts of its territory at other acceptance areas in the cone-shaped system. The adoption of the during temporal primitive for the characterization of North, East, South and West directions was motivated by the assumption that the *centroid* of the primary object is located in the zone of acceptance for those directions, as defined by the reference object.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

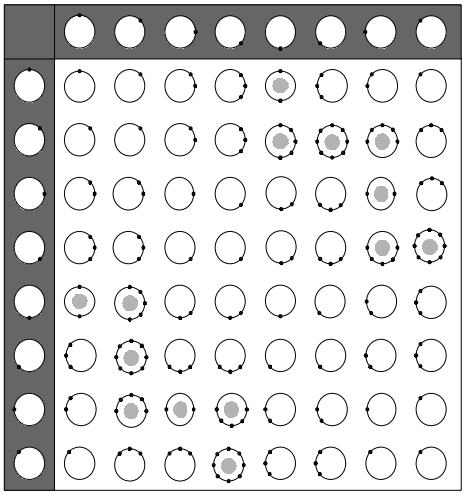
Figure 14. Temporal Intervals for the Characterization of Direction and Topology for Administrative Subdivisions



In the case of adjacency it is clear by an analysis of *Figure 14* that some overlapping between the regions can exist, when analyzed in a temporal perspective. This fact influenced the adoption of the overlap and overlapped by primitives instead of the meet and met by primitives adopted by Sharma.

Following the assumptions described above new composition tables were constructed. *Table 4* shows the particular case of integration of direction with the topological pair disjoint;disjoint. The other composition tables, for the topological pairs disjoint;meet, meet;disjoint and meet;meet, are available in Santos (2001).

Table 4. Composition Table for the Integration of Direction with the Topological Pair disjoint;disjoint (particular case of administrative subdivisions)



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

After the identification of the composition tables that integrate direction and topology under the principles of the cone-shaped system, it was possible to integrate these tables with the composition table proposed by Hong (1994), with respect to direction and distance. This step was preceded by a detailed analysis of the application domain in which the system will be used, composition of regions that represent administrative subdivisions that cover all the territory considered, without any gap or overlap (Santos, 2001). Concerning to the distance spatial relation, it was defined that the qualitative distance very close is restricted to adjacent regions. When the qualitative distance is close the regions may be, or may not be, adjacent. The far and very far qualitative distances can only exist between regions that are disjoint from each other.

The basic assumption for the integration process was that the outcome direction in the integration of direction and distance is the same outcome direction in the integration of direction and topology, or it belongs to the set of possible directions inferred by the last one. The direction that guides the integration process is the direction suggested by the composition table of direction and distance (it is more accurate since it considers the distance existing between the objects).

The final composition table, which is shown with the graphical symbols expressed in *Figure 15*, was obtained through an integration process that is diagrammatically demonstrated in *Figure 16*. For example, the composition of (North, very close) with (North, very close) has as result (North, very close). The composition of (North, meet) with (North, meet) has as the result (North, disjoint or meet). The integration of the three spatial relations leads to (North, very close, disjoint or meet). As the qualitative distance relation very close was restricted to adjacent regions, the result of the integration is (North, very close, meet). Another example explicit in *Figure 16* is the integration of the result of (North, close);(North, close) with (North, disjoint);(North, disjoint). The result of the first composition is (North, far) while the result of the second is (North, disjoint). The integration generates the value (North, far, disjoint), which matches the principles adopted in this work for the distance relation: if the regions are far from each other, then topologically they are disjoint.

In the evaluation of the composition table constructed it was realized that the dimensions of the regions influenced (sometimes negatively) the results achieved. Qualitative reasoning with administrative subdivisions is a difficult task, which is influenced not only by the irregular limits of the regions but also by their size. As can be noted in *Figure 17*[9], if the dimension of A is lower then the dimension of B, and the dimension of B is lower than the dimension of C, then the inference result must be A Northeast C. But if the dimension of A is greater than the dimension of B and the dimension of B is lower than the dimension of C, then the inference result must be A North C. A detailed analysis of

Figure 15. Graphical Representation of Direction, Distance and Topological Spatial Relations

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 16. Integration of Direction, Distance and Topological Spatial Relations

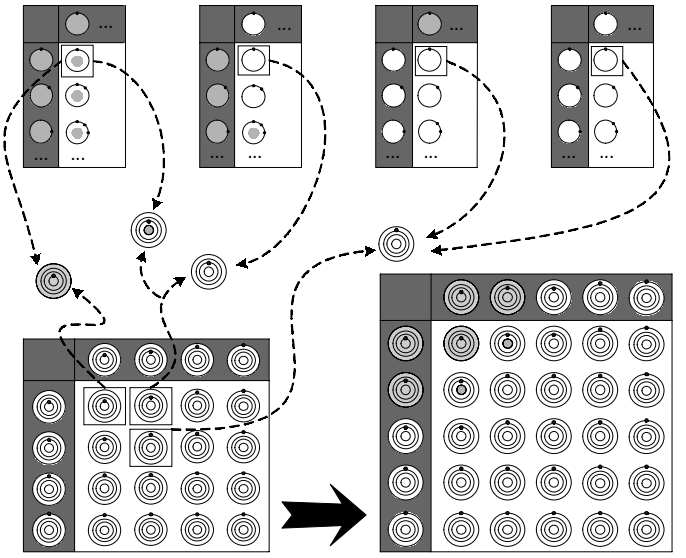
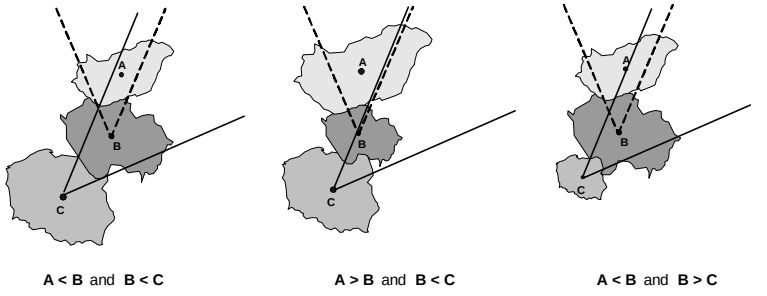


Figure 17. Influence of the Regions Dimension in the Inference Result



these situations was undertaken, allowing the identification of several rules that integrate the dimensions of the regions in the qualitative reasoning process of the PADRÃO system. Through this process, the reasoning process was improved, and more accurate inferences were obtained.

The performance of the qualitative reasoning system was evaluated (Santos, 2001). The approach followed in this performance test was to compare the spatial relations obtained through the qualitative inference process with the spatial relations obtained by quantitative methods. A Visual Basic module was implemented for the execution of this task. This module calculated quantitatively all the spatial relations existing between the Municipalities of three districts of Portugal, looking at the position of the respective *centroids*. This information was stored in a table and compared with the spatial relations inferred

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

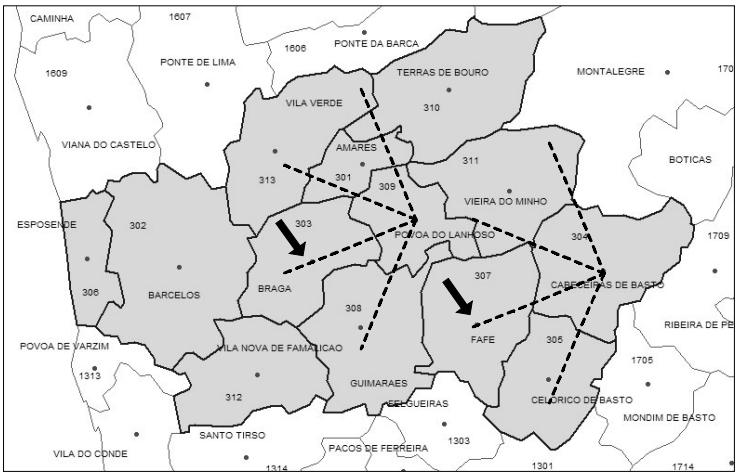
qualitatively. The results achieved were, in the poor scenario, exact[10] for 75% of the inferences obtained in Districts with higher differences between the dimensions of their regions (two of the analyzed Districts). For the Braga District, a District that integrates regions with homogeneous dimensions, the inferences obtained were 88% exact for direction and 81% exact for distance. For topology, the inferences were in all cases 100% exact. The approximate inferences obtained were verified in regions that have parts of their territory in more than one acceptance area for the direction relation. For these cases, the *centroid* of the region is sometimes positioned in one acceptance area, although the region has parts of its territory in other acceptance areas. Another situation, as shown in *Figure 18* for two Municipalities, is verified when the *centroid* is positioned in the line that divides the acceptance areas, which makes even more difficult the identification of the direction between the regions and, as a consequence, the qualitative reasoning process.

After the evaluation of the qualitative reasoning system implemented and the analysis of the inferences obtained, which provided a good approximation to the reality, the system will be afterwards used in the knowledge discovery process.

The PADRÃO System

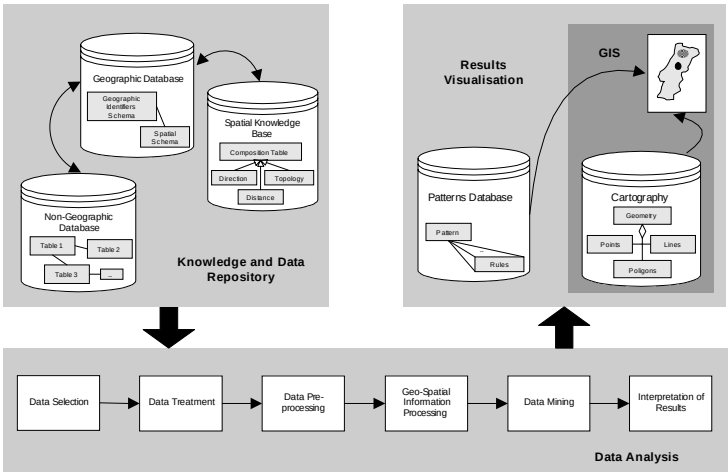
PADRÃO is a system for knowledge discovery in geo-referenced databases based on qualitative spatial reasoning. This section presents its architecture, gives some technical details about its implementation and tests the system in a geo-referenced data set.

Figure 18. Municipalities of the Braga District



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 19. Architecture of PADRÃO



Architecture of PADRÃO

The architecture of PADRÃO (*Figure 19*) aggregates three main components: Knowledge and Data Repository, Data Analysis and Results Visualization. The Knowledge and Data Repository component stores the data and knowledge needed in the knowledge discovery process. This process is implemented in the Data Analysis component, which allows for the discovery of patterns or other relationships implicit in the analyzed geo-spatial and non-spatial data. The discovered patterns can be visualized in a map using the Results Visualization component. These components are described below.

The Knowledge and Data Repository component integrates three central databases:

1. A *Geographic Database* (GDB) constructed under the principles established by the European Committee for Normalization in the CEN TC 287 pre-standard for Geographic Information. Following the pre-standard recommendations it was possible to implement a GDB in which the positional aspects of geographic data are provided by a *geographic identifiers system* (CEN/TC-287, 1998). This system characterizes the administrative subdivisions of Portugal at the municipality and district level. Also it includes a geographic gazetteer containing the several geographic identifiers used and the concept hierarchies existing between them. The geographic identifiers system was integrated with a *spatial schema* (CEN/TC-287, 1996) allowing for the definition of the *direction*, *distance* and *topological* spatial relations that exist between adjacent regions at the Municipality level.
2. A *Spatial Knowledge Base* (SKB) that stores the qualitative rules needed in the inference of new spatial relations. The knowledge available in this database aggregates the constructed composition table (integrating direction, distance and topological spatial relations), the set of identifiers used, and the several rules that incorporate the dimension of the regions in the reasoning process. This knowledge

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

base is used in conjunction with the GDB in the inference of unknown spatial relations.

3. A *non-Geographic Database* (nGDB) that is integrated with the GDB and analyzed in the Data Analysis component. This procedure enables the discovery of implicit relationships that exist between the geo-spatial and non-spatial data analyzed.

The Data Analysis component is characterized by six main steps. The five steps presented above for the knowledge discovery process plus the Geo-Spatial Information Processing step. This step verifies if the geo-spatial information needed is available in the GDB. In many situations the spatial relations are implicit due to the properties of the spatial schema implemented. In those cases, and to ensure that all geo-spatial knowledge is available for the data mining algorithms, the implicit relations are transformed into explicit relations through the inference rules stored in the SKB.

The Results Visualization component is responsible for the management of the discovered patterns and their visualization in a map (if required by the user and when the geometry[11] of the analyzed region is available). For that PADRÃO uses a Geographic Information System (GIS), which integrates the discovered patterns with the geometry of the region. This component aggregates two main databases:

1. The *Patterns Database* (PDB) that stores all relevant discoveries. In this database each discovery is catalogued and associated with the set of rules that represents the discoveries made in a given data mining task.
2. A *Cartographic Database* (CDB) containing the cartography of the region. It aggregates a set of points, lines and polygons with the geometry of the geographical objects.

Implementation of PADRÃO

PADRÃO was implemented using the relational database system Microsoft Access, the knowledge discovery tool Clementine (SPSS, 1999), and Geomedia Professional (Intergraph, 1999), the GIS used for the graphical representation of results.

The databases that integrate the Knowledge and Data Repository and the Results Visualization components were implemented in Access. The data stored in them are available to the Data Analysis component or from it, through ODBC (Open Database Connectivity) connections.

Clementine is a data mining toolkit based on visual programming[12], which includes machine learning technologies like rule induction, neural networks, association rules discovery and clustering. The knowledge discovery process is defined in Clementine through the construction of a *stream* in which each operation on data is represented by a *node*.

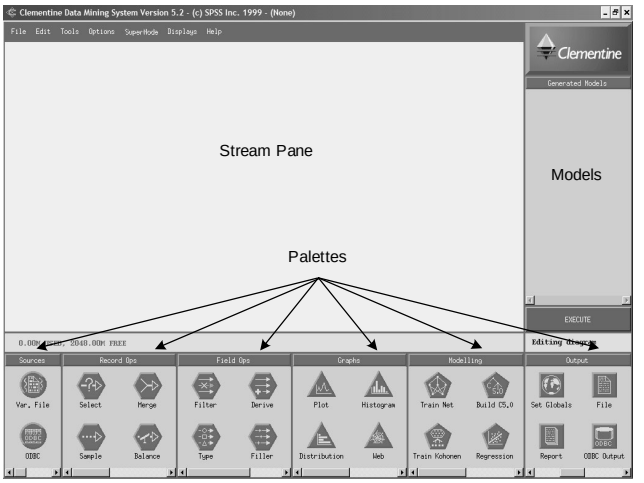
The workspace of Clementine comprises three main areas. The main work area, the Stream Pane, constitutes the area for the streams construction. The palettes area in which the several available icons are grouped according to their functions: links to sources of information, operations on data (rows or columns), visual facilities and modeling

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

techniques (data mining algorithms). The models area stores the several models generated in a specific stream. These models can be directly re-used in other streams or they can be saved providing for their later use. *Figure 20* shows the work environment of Clementine and presents some of the several nodes available according to their functionality. Circular nodes represent links to data sources and constitute the first node of any stream. Nodes with a hexagonal shape are for data manipulation, including operations on records (lines of a table) or operations on fields (columns of a table). Triangular nodes allow for data exploration and visualization, providing a set of graphs that can be used to get a better understanding of data. Nodes with a pentagonal shape are modeling nodes, i.e., data mining algorithms that can be used to identify patterns in data. The last group of square-shaped nodes is related to the output functions, which make available a set of nodes for reporting, storing or exporting data.

The Data Analysis component of PADRÃO is based on the construction of several streams that implement the knowledge discovery process. The several models obtained in the data mining phase represent knowledge about the analyzed data and can be saved or reused in other streams. In PADRÃO, these models can be exported through an ODBC connection to the PDB. The integration of the PDB with the CDB allows the visualization of the rules explicit in the models in a map. The visualization is achieved through the VisualPadrão application, a module implemented in Visual Basic. VisualPadrão manipulates the library of objects available in Geomedia. This application was integrated in the Clementine workspace using a *specification file*, i.e., a mechanism provided by the Clementine system that allows for the integration of new capabilities in its environment. This approach provides an integrated workspace in which all tasks associated with the knowledge discovery process can be executed.

Figure 20. Workspace of Clementine



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Analysis of a Geo-Referenced Database

Several datasets have been analyzed by the PADRÃO system. Among them are demographic databases storing the Parish records of several Municipalities of Portugal (Santos & Amaral, 2000a; Santos & Amaral, 2000b; Santos & Amaral, 2000c). Another dataset analyzed was a component of the Portuguese Army Database (Santos & Ramos, 2003). The several data mining objectives defined allowed for the identification of the implicit relationships existing between the geo-spatial and non-spatial data analyzed.

The dataset selected for description in this chapter integrates data from a financial institution, which supplies credit for the acquisition of several types of goods. To overcome confidentiality issues with the data and the several identifiable patterns, the data was manipulated in order to create a random data set. Through this process the confidentially is ensured and the knowledge discovery process in the PADRÃO system can be described.

The bank database aggregates a set of 3,031 records that characterize the behavior of the bank clients. For this data a data mining objective, "*identify the profile of the clients in order to minimize the institutional risk of investment*" was defined. This profile will be identified for the Braga District, one of the Districts of Portugal.

The knowledge discovery process is preceded by the business understanding phase in which the meaning and importance of each attribute for this process is evaluated. The attributes integrated in the database are: identification number (ID), VAT number (VAT_number), client title (Title), name (Name), good purchased (Acquisition), contract duration (Duration), income (Salary), overall value of credit (Credit_value), payment type (Payment_type), credit for home acquisition (Home_credit), lending value (Payment_value), marital state (Marital_state), number of children (Number_child), age (Age) and the accomplishment or not of the credit (Fault).

At this phase Distribution and Histogram nodes of Clementine were used to explore the several attributes, identify their values, identify their distribution, and determine if any of them present anomalies. *Figure 21* shows the stream constructed for this exploration phase.

The results obtained by each Distribution[13] graph are showed in *Figure 22*. It can be seen that the majority of attributes present a distribution of values that are the normal operation of the organization. However, exceptions were verified for the Home_credit and Title attributes. Namely:

- The attribute Home_credit, which shows if the client has or does not have a credit for home acquisition (values 1 and 0 respectively), also includes a record with the 2 value. As this value constitute an error, the respective record must be removed from the dataset;
- The Title attribute integrates five cases of credit for organizations (value Company). As a result, these records must be removed from the database as they represent a minority class[14] in the overall set.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 21. Data Exploration

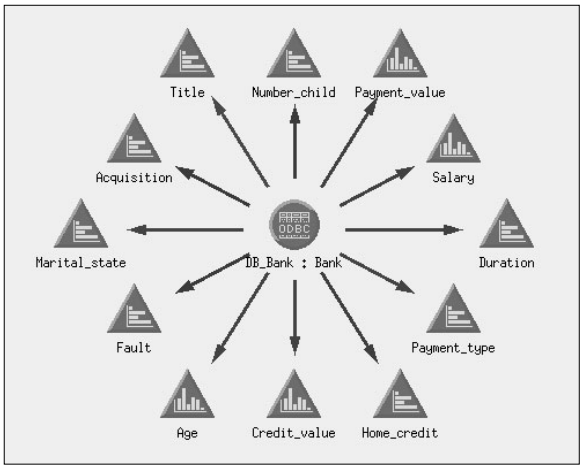


Figure 22. Distribution of Categorical Data

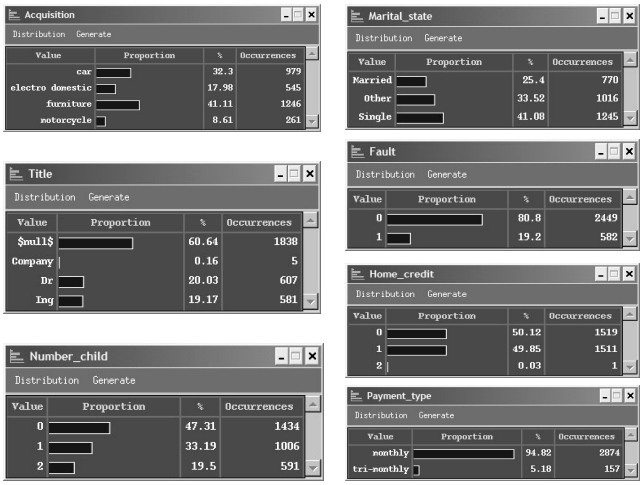


Figure 23 shows the Histograms with the distribution of attributes with continuous values. The analysis of the distributions allows for the verification of the several classes that will be created in order to transform continuous values into discreet values. The defined classes are presented in *Table 5*. Their definition is based on the assumption that the data available for analysis must be distributed homogeneously across the several classes.

This exercise of exploration and comprehension of the available data allowed the identification of the attributes for analysis and the definition of the several classes that will be used in the pre-processing step, i.e., to transform continuous values into discreet

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 23. Distribution of Continuous Values

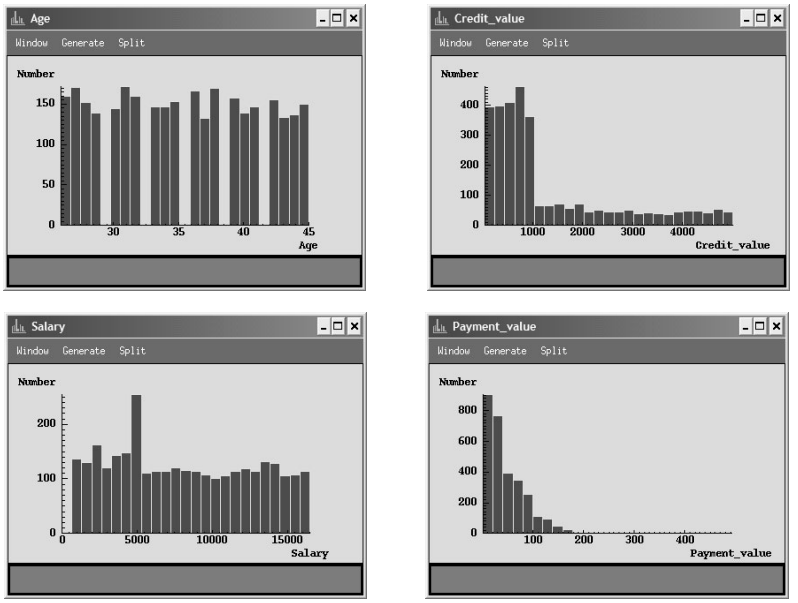


Table 5. Classes for Attributes with Continuous Values

Attributes	Classes
Age	(25..31] → '**26-31**', (31..38] → '**32-38**', (38..45] → '**39-45**'
Credit_value	(0..350] → '**0-350**', (350..650] → '**351-650**', (650..900] → '**651-900**', (900..2500] → '**901-2500**', (2500..5000] → '**2501-5000**'
Salary	(0..4500] → '**0-4500**', (4500..8000] → '**4501-8000**', (8000..12500] → '**8001-12500**', (12500..17000] → '**12501-17000**'
Payment_value	(0..17] → '**0-17**', (17..30] → '**18-30**', (30..50] → '**31-50**', (50..80] → '**51-80**', (80..500] → '**81-500**'

values. Next, the six steps considered in the PADRÃO system for the knowledge discovery process (Data Selection, Data Treatment, Data Pre-processing, Geo-spatial Information Processing, Data Mining and Interpretation of Results) are described.

Data Selection and Data Treatment

The data selection step allows for the exclusion of attributes that have no influence in the knowledge discovery process. Among them are ID, VAT_number, Title and Name, since they only have an informative role. The other attributes will be considered in order to evaluate the contribution of each one to the definition of the profile of the clients.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 24. Data Selection and Data Treatment Steps

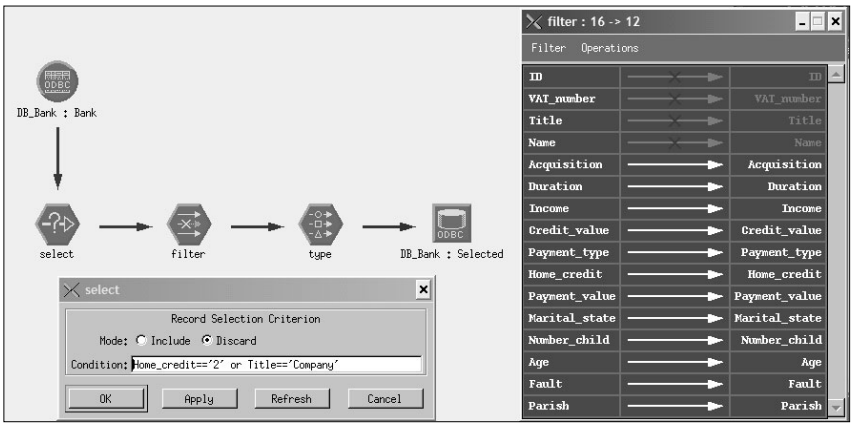


Figure 24 shows the stream constructed for the data selection and data treatment steps. The stream integrates a source node (DB_Bank:Bank) that makes the data available to the knowledge discovery process through an ODBC connection. The select node discards records with anomalies. As previously mentioned, the record with the value 2 in the Home_credit attribute must be deleted. All records associated with the value Company in the Title attribute need to be also removed. The filter node is used to select the attributes that will be excluded from the process. The type node allows for the specification of the data type (numeric, character …) of the attributes that will be exported to the database. As result of the several tasks undertaken, a new table (DB_Bank:SelectedData) is created in the bank database.

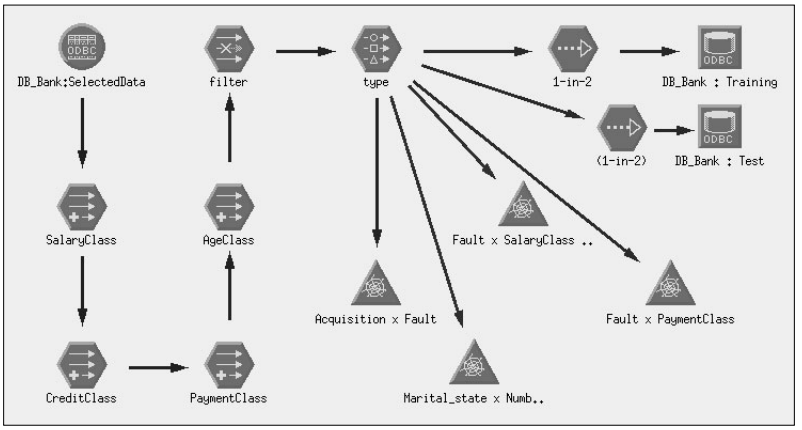
Data Pre-Processing

The data pre-processing step (*Figure 25*) allows for the transformation of the attributes with continuous values into attributes with discreet values (nodes SalaryClass, CreditClass, PaymentClass and AgeClass), according to the classes presented in *Table 5*. In this step, web nodes, exploration graphs available in Clementine, are also used for the identification of associations[15] among the analyzed attributes (nodes Acquisition x Fault, SalaryClass x AgeClass x Fault, Marital_state x Number_child x Fault and PaymentClass x Fault). The last task undertaken is associated with the creation of the two datasets (nodes DB_Bank:Training and DB_Bank:Test) that will be used from now on. They are the Training and the Test datasets, and in which the original data is randomly distributed. The Training file is used in the model construction (data mining step) while the Test dataset evaluates the model confidence when applied to unknown data.

The web nodes constructed are shown in *Figure 26*. They combine several attributes and through the analysis of them it is possible to identify associations between attributes. Strong associations between attributes are represented by bold lines, while weak

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 25. Data Pre-Processing Step



associations are symbolized by dotted lines. For the several acquisitions that can be effected, *Figure 26a* points out that no association exists between the good furniture and the value 1 of the Fault attribute, indicating that faults were not usual with the credit supplied for this specific acquisition. Analyzing the income and age attributes with Fault in *Figure 26b*, it is evident that individuals with a higher income honor their payments, since the value 12501-17000 of the SalaryClass attribute presents no association with value 1 of Fault. Between value 8001-12500 and value 1 of Fault there exists a weak association, which indicates that this specific group may or may not be able to honor its credit payments. Similarly, a weak association is verified between the marital state Single and value 1 of Fault, *Figure 26c*. PaymentClass and Fault present strong connections between all attribute values as seen in *Figure 26d*, thus indicating that all type of payment values are associated with *good* and *bad* clients.

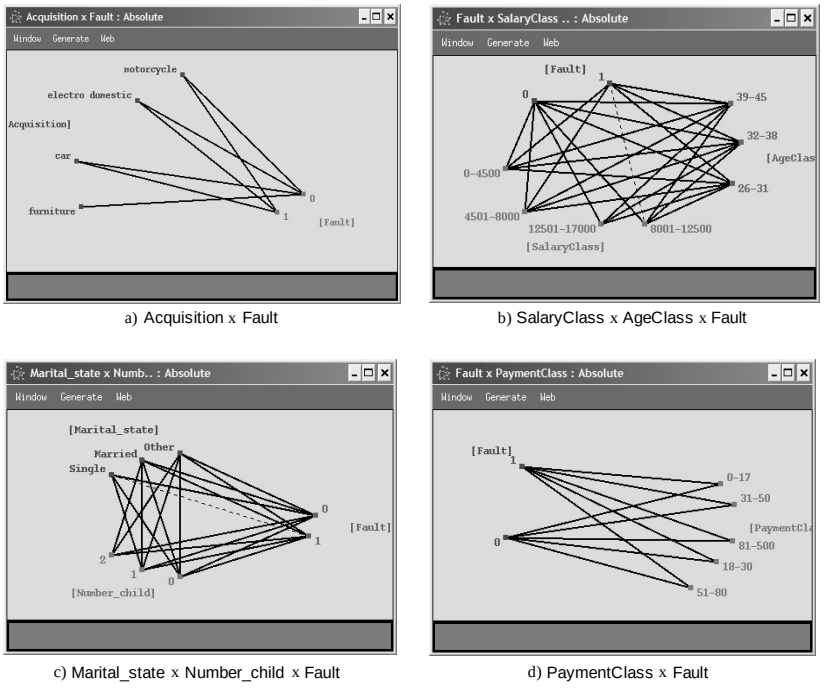
Geo-Spatial Information Processing

As the GDB only stores spatial relations for adjacent regions and, as it is necessary to analyze if the geographical component has any influence in the identification of the profile of the clients, all the other relationships that exist between non-adjacent regions and needed in the data mining step will be inferred. In Clementine, a rule induction[16] algorithm is able to learn the inference rules available in the composition table stored in the SKB. That enables the inference of new spatial relations.

The models created, nodes infDir, infDis and infTop, can now be used in the inference process. With these models and as shown in *Figure 27* it is possible to infer the unknown spatial relationships existing in the Municipalities of the Braga District. The spatial relations for adjacent regions stored in the GBD are gathered through the source node (GDB:geoBraga) of the stream and combined (node Inflection) in order to obtain new associations between regions. The spatial relations existing among these new associations are identified by the models infDir, infDis and infTop. After the inferential process,

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 26. Data Exploration with Web Nodes



a) Acquisition x Fault

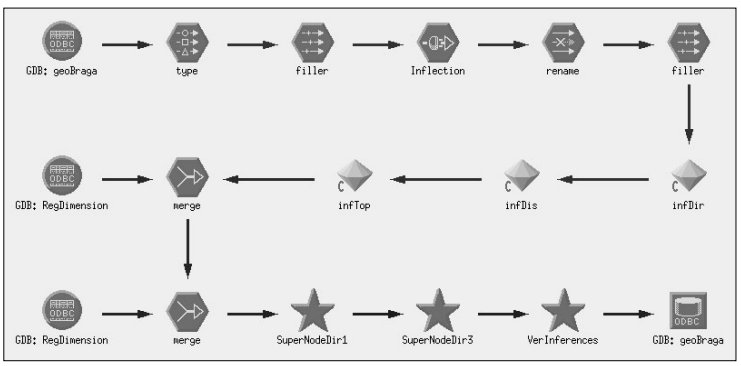
b) SalaryClass x AgeClass x Fault

c) Marital_state x Number_child x Fault

d) PaymentClass x Fault

the knowledge obtained is recorded in the GDB (output node GDB:geoBraga). In the stream of *Figure 27*, the super nodes SuperNodeDir1 and SuperNodeDir3 are responsible for the integration of the dimension of the regions in the reasoning process. In this process, there is validation if the several inferences obtained for a particular region agree independently of the composed regions. Several paths can be followed in order to infer a specific spatial relation. For example, knowing the facts A North B, A East D, B East C and

Figure 27. Geo-Spatial Information Processing Step



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

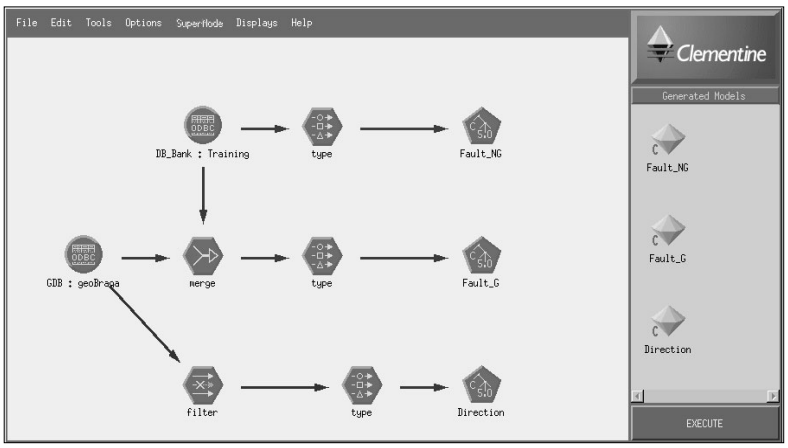
D North C, the direction relation existing between A and C may be obtained composing A North B with B East C or combining A East D with D North C. If several compositions can be effected and if the results obtained from each one do not match, then the super node VerInferences excludes those results from the set of accepted ones.

Data Mining

In the data mining step (*Figure 28*) an appropriate algorithm is selected to carry out a specific data mining task. Three different tasks were undertaken (see *Figure 28*). First, a decision tree (node Fault_NG) that characterizes the profile of the clients without considering the location of the clients was generated. Second, the training set (DB_Bank:Training) was integrated with the spatial relations for the District in analysis (GDB:GeoBraga) in order to include the geographical component in the analysis of the profile of the clients (node Fault_G). Third, the geographical model of the District was created. This latter model (Direction) indicates the direction of each Municipality in the District and was obtained by analyzing the spatial relations inferred in the geo-spatial information processing step. All models were obtained with the C5.0 algorithm that allows for the induction of decision trees. *Figure 28* highlights the stream constructed for the generation of the three models. These models are available in the Generated Models palette and have the shape of a diamond (right hand side of *Figure 28*).

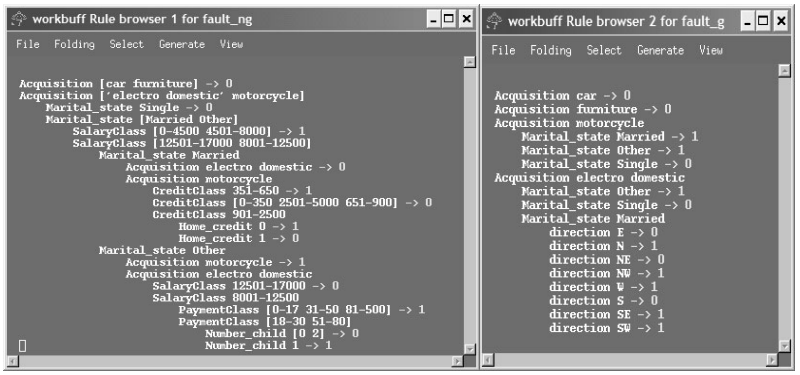
The Fault_NG model (*Figure 29*, left side) integrates a set of rules that are represented in a decision tree, which characterizes the profile of the clients. Through the analysis of the model it is possible to verify that the acquisition of car and furniture is traditionally associated with clients that honor their payments, while the acquisition of electro domestic and motorcycle have other attributes (Marital_state, Salary_class ...) that influence the profile of the clients. One explicit rule in the model for clients that the institution has no interest in supplying with credit, is: IF SalaryClass = '12501-17000' and Marital_state= 'Married' and Acquisition = 'motorcycle' and CreditClass = '351-600' THEN 1. The Fault_G model

Figure 28. Data Mining Step



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 29. Generated Models for the Profile of the Clients

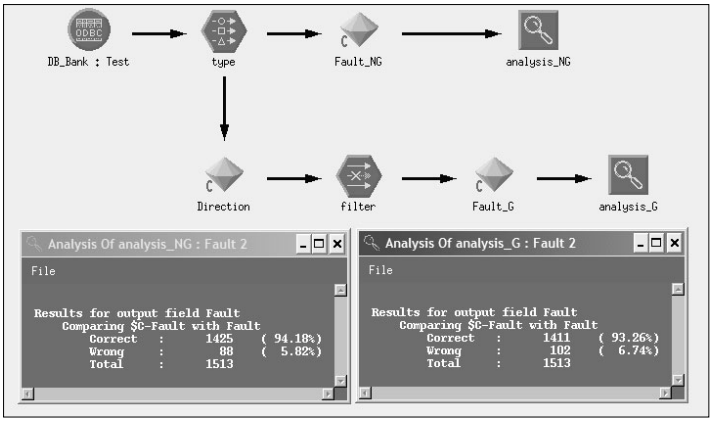


(*Figure 29*, right hand side) allows for the verification of geographic zones that have associated clients with a higher incidence of faults. These zones are represented in directions, which partition the District into eight areas. The analysis of the model points out that Northeast (NE), East (E) and South (S) are associated with clients that pay the credit assumed.

Interpretation of Results

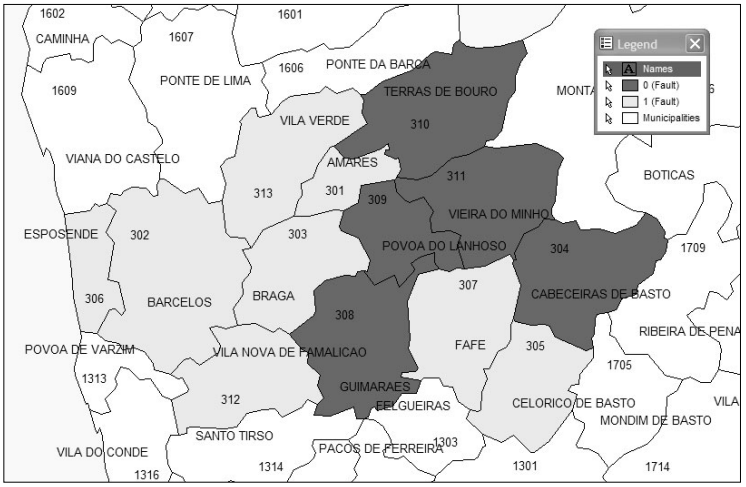
The Test set (DB_Bank:Test) is used in the interpretation of results step to verify the confidence of the models built in the Data Mining step. With respect to the Fault_NG model, *Figure 30* shows a percentage of confidence of 94.18%. The Fault_G model presents a percentage of confidence of 93.26%. This decrease in the model confidence, when considering the geographical component, may be caused by the aggregation of

Figure 30. Percentage of Confidence of the Generated Models



Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 31. Visualization of Results



Municipalities into eight regions (the Cardinal directions), which represents a loss of specificity in favor of generality. Although the Direction model was obtained through the analysis of spatial relations inferred by qualitative rules, the results obtained in the Fault_G model maintain a high level of confidence.

The PADRÃO system permits the visualization of the results of the knowledge discovery process on a map. In this system the several rules that integrate a model are recorded in the PDB (Santos & Amaral, 2000b). At the same time the user has the option to run the VisualPadrão tool and visualizes the desired model (*Figure 31*). As can be noted in the figure, Municipalities located at Northeast, East and South of the District contain clients mainly associated with no faults in their credit payments (information explicit in the Fault_G model obtained in the data mining step). This geographic characterization enabled the identification of regions where the relative incidence of clients with faults is higher than elsewhere in the District.

Risk zones were identified, aggregating together regions that have clients with similar behavior. The geographic segments can be cataloged by the bank, looking at similarities like proximity with other regions, population density, population qualification and other relevant issues.

The models obtained in the data mining step define the profile of the bank clients. They integrate the attributes and the corresponding values related to the classification of the clients bearing in mind the risk of investment in specific classes of clients. For the available segments, the several rules identified can support managers in the decision-making process. In the granting of new credits, the organization is now supported by models that track the previous behavior of its clients, indicating groups of clients in which the organization has to pay more attention in the granting of credit and those groups without difficulties in the assignment of a credit.

Suppose that 10 new potential clients request credit to the organization. *Figure 32* shows the relevant data on each client and the classification (column $C-Fault) of the model

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Figure 32. Classification of New Clients by the Model

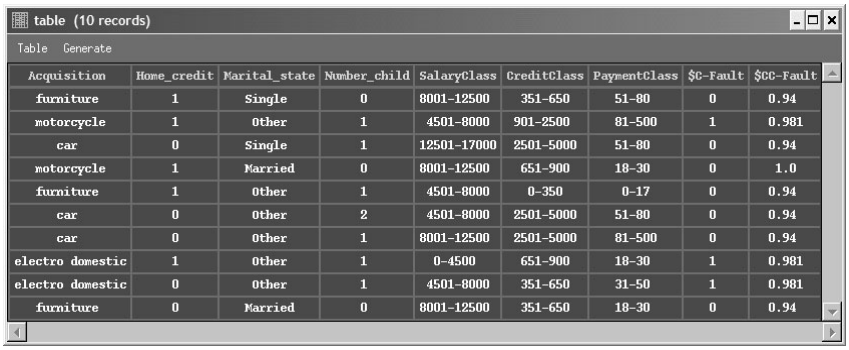


table (10 records)

Table Generate

Acquisition	Home_credit	Marital_state	Number_child	SalaryClass	CreditClass	PaymentClass	$C-Fault	$CC-Fault
furniture	1	Single	0	8001-12500	351-650	51-80	0	0.94
motorcycle	1	Other	1	4501-8000	901-2500	81-500	1	0.981
car	0	Single	1	12501-17000	2501-5000	51-80	0	0.94
motorcycle	1	Married	0	8001-12500	651-900	18-30	0	1.0
furniture	1	Other	1	4501-8000	0-350	0-17	0	0.94
car	0	Other	2	4501-8000	2501-5000	51-80	0	0.94
car	0	Other	1	8001-12500	2501-5000	81-500	0	0.94
electro domestic	1	Other	1	0-4500	651-900	18-30	1	0.981
electro domestic	0	Other	1	4501-8000	351-650	31-50	1	0.981
furniture	0	Married	0	8001-12500	351-650	18-30	0	0.94

according to the rules explicit in it. The column $CC-Fault indicates the confidence of the classification, which is equal or superior to 94%. Looking at the classification achieved, for seven clients the decision of the model is 0 in the $C-Fault attribute, which means that based in the past experience of the organization these are *good* clients. For 3 clients the result was 1 in the $C-Fault attribute, labeling these clients as *risk* clients and suggesting that a more detailed analysis must be undertaken in order to identify the appropriate decision (grant credit or not).

The use of predictive models assumes that the past is a good predictor of the future. However, there are situations where the past may not be a good predictor, if the facts occurred were influenced by external events not present in the analyzed data (Berry & Linoff, 2000). For this reason, and when talking about prediction, the organization cannot only be supported by the models obtained in the knowledge discovery process, in order to avoid the penalization of new potential *good* clients as a result of the behavior of past clients. The models obtained should be seen as tools that support the decision-making process, not as the decision-maker.

The knowledge discovery process should support the creation of organizational knowledge through the incorporation of the information expressed in the several models in its daily activities. This procedure will contribute to fulfill the information requirements of the bank and help in the accomplishment and improvement of its mission.

Conclusion

This chapter presented an approach for knowledge discovery in geo-referenced databases based on qualitative spatial reasoning, where the position of geographical data was provided by qualitative identifiers.

Direction, distance and topological spatial relations were defined for a set of Municipalities of Portugal. This knowledge and the composition table constructed for integrated spatial reasoning, about direction, distance and topological relations, allowed for the

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

inference of new spatial relations analyzed in the data mining step of the knowledge discovery process.

The integration of a bank database with the GDB (storing the administrative subdivisions of Portugal) made possible the discovery of general descriptions that exploit the relationships that exist between the geo-spatial and non-spatial data analyzed. The models obtained in the data mining step define the profile of the clients, bearing in mind the risk of investment of the organization for specific segments of clients. For the available classes, the several rules identified support the managers of the organization in the decision-making process. The latter represents one of the organizational processes that can benefit from data mining technology through the incorporation of its results in the evaluation of critical and uncertain situations.

The results obtained with the PADRÃO system point out that traditional KDD systems, which were developed for the analysis of relational databases and that do not have semantic knowledge linked to spatial data, can be used in the process of knowledge discovery in geo-referenced databases, since some of this semantic knowledge and the principles of qualitative spatial reasoning are available as domain knowledge. Clementine, a KDD system, was used in the assimilation of the geographic domain knowledge such as composition tables, in the inference of new spatial relations, and in the spatial patterns discovery.

The main advantages of the proposed approach, for mining geo-referenced databases, include the use of already existing data mining algorithms developed for the analysis of non-spatial data; an avoidance of the geometric characterization of spatial objects for the knowledge discovery process; and the ability of data mining algorithms to deal with geo-spatial and non-spatial data simultaneously, thus imposing no limits and constraints on the results achieved.

Acknowledgments

We thank NTech – Sistemas de Informação, Lda. for making the database available for analysis. We thank Tony Lavender for his help in improving the English writing of this chapter.

References

Abdelmoty, A. I., & El-Geresy, B. A. (1995). A general method for spatial reasoning in spatial databases. *Proceedings of the Fourth International Conference on Information and Knowledge Management* (pp. 312-317). Baltimore, Maryland.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM, 26* (11), 832-843.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Berry, M., & Linoff, G. (2000). *Mastering data mining: The art and science of customer relationship management*. New York: John Wiley & Sons.

CEN/TC-287. (1996). *Geographic information: Data description, spatial schema* (prENV 12160). European Committee for Standardization.

CEN/TC-287. (1998). *Geographic information: Referencing, geographic identifiers* (prENV 12661). European Committee for Standardization.

Egenhofer, M. J. (1994). Deriving the composition of binary topological relations. *Journal of Visual Languages and Computing, 5* (2), 133-149.

Ester, M., Kriegel, H.-P., & Sander, J. (1997). Spatial data mining: A database approach. *Proceedings of the Fifth International Symposium on Large Spatial Databases* (pp. 47-68). Germany.

Ester, M., Frommelt, A., Kriegel, H.-P., & Sander, J. (1998). Algorithms for characterization and trend detection in spatial databases. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press.

Fayyad, U., & Uthurusamy, R. (1996). Data mining and knowledge discovery in databases. *Communications of the ACM, 39* (11), 24-26.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (eds.). (1996). *Advances in knowledge discovery and data mining*. MA: The MIT Press.

Frank, A. U. (1992). Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages and Computing, 3*, 343-371.

Frank, A. U. (1996). Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Systems, 10* (3), 269-290.

Freksa, C. (1992). Using orientation information for qualitative spatial reasoning. In A. U. Frank, I. Campari, & U. Formentini (Eds.), *Theories and methods of spatio-temporal reasoning in geographic space* (Lectures Notes in Computer Science 639). Berlin: Springer-Verlag.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. CA: Morgan Kaufmann Publishers.

Han, J., Tung, A., & He, J. (2001). SPARC: Spatial association rule-based classification. In R. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar & R. Namburu (Eds.), *Data mining for scientific and engineering applications* (pp. 461-485). Kluwer Academic Publishers.

Hernández, D., Clementini, E., & Felice, P. D. (1995). Qualitative distances. *Proceedings of the International Conference COSIT'95* (pp. 45-57). Austria.

Hong, J.-H. (1994). *Qualitative distance and direction reasoning in geographic space*. Unpublished doctoral dissertation, University of Maine, Maine.

Intergraph. (1999). *Geomedia professional v3* (Reference Manual). Intergraph Corporation.

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information systems. *Proceedings of the 4th International Symposium on Large Spatial Databases* (pp. 47-66). Maine.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Lu, W., Han, J., & Ooi, B. (1993). Discovery of general knowledge in large spatial databases. *Proceedings of the 1993 Far East Workshop on Geographic Information Systems* (pp. 275-289). Singapore.

Papadias, D., & Sellis, T. (1994). On the qualitative representation of spatial knowledge in 2D space. *Very Large Databases Journal, Special Issue on Spatial Databases, 3* (4), 479-516.

Santos, M., & Amaral, L. (2000a). Knowledge discovery in spatial databases through qualitative spatial reasoning. *Proceedings of the 4th International Conference and Exhibition on Practical Applications of Knowledge Discovery and Data Mining* (pp. 73-88). Manchester.

Santos, M., & Amaral, L. (2000b, November). Knowledge discovery in spatial databases: The PADRÃO's qualitative approach. *Cities and Regions, GIS special issue*, 33-49.

Santos, M., & Amaral, L. (2000c). A qualitative spatial reasoning approach in knowledge discovery in spatial databases. *Proceedings of Data Mining 2000: Data Mining Methods and Databases for Engineering, Finance and Others Fields* (pp. 249-258). Cambridge.

Santos, M. Y. (2001). *PADRÃO: Um sistema de descoberta de conhecimento em bases de dados geo-referenciadas* (*in Portuguese*). Unpublished doctoral dissertation, Universidade do Minho, Portugal.

Santos, M. Y., & Ramos, I. (2003). Knowledge construction: The role of data mining tools. *Proceedings of the UKAIS 2003 Conference "Co-ordination and Co-operation: the IS role."* Warwick.

Sharma, J. (1996). *Integrated spatial reasoning in geographic information systems: Combining topology and direction*. Unpublished doctoral dissertation, University of Maine, USA.

SPSS. (1999). *Clementine* (user guide, version 5.2). SPSS Inc.

Endnotes

[1] GISs allow for the storage of geographic information and enable users to request information about geographic phenomena. If the requested spatial relation is not explicitly stored in databases, it must be inferred from the information available. The inference process requires searching relations that can form an *inference path* between the two objects where the relation is requested (Hong, 1994). The composition operation combines two contiguous paths in order to infer a third spatial relation. A composition table integrates a set of inference rules used to identify the result of a specific composition operation.

[2] Extended objects are not point-like, so represent objects for which their dimension is relevant (Frank, 1996). In this work, extended objects are geometrically represented by a polygon, indicating that their position and extension in space are relevant.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

[3] In IR^2, there are eight topological relations between two planar regions without holes (two-dimensional, connected objects with connected boundaries); 18 topological relations between spatial regions with holes; 33 between two simple lines and 19 between a spatial region without holes and a simple line (Egenhofer, 1994).

[4] The topology of a full planar graph refers to a planar graph that integrates regions completely covering the plane without any gap or overlap. Regions are topologically represented by faces, which are defined without holes (CEN/TC-287, 1996).

[5] Defining distances between regions is a complex task, since the size of each object plays an important role in determining the possible distances. Sharma (Sharma, 1996) enumerates as possible ways to the definition of distances between regions: (i) taking the distance between the *centroids* of the two regions; (ii) determining the shortest distance between the two regions; or (iii) determining the furthest distance between the two regions.

[6] Other validity intervals, for different ratios, can by found in Hong (1994).

[7] The symbol used to represent the composition operation is ";".

[8] Since the system will be used with administrative subdivisions, the orientation between the several regions is calculated according to the position of the respective *centroids*.

[9] The dotted lines define the acceptance area defined for the North direction (designed from the *centroid* of B), while the whole lines represent the acceptance area defined for the Northeast direction (designed from the *centroid* of C).

[10] In this work, an inference is considered exact if the result achieved with the correspondent qualitative rule is the same as if the data was translated to quantitative information and manipulated through analytical functions. Otherwise, it is considered approximate.

[11] The geometry is not required in the knowledge discovery process, since the manipulation of the geographic information is undertaken by a qualitative approach (as described in previous sections).

[12] Visual programming involves placing and manipulating icons representing processing nodes.

[13] Distribution nodes are used for the analysis of categorical data.

[14] The data mining algorithms may be negatively influenced by classes with a great number of values.

[15] The several associations identified anticipate the importance of each attribute in the definition of the profile of the clients.

[16] A rule induction algorithm creates a decision tree aggregating a set of rules for classifying the data into different outcomes. This technique only includes in the rules the attributes that really matter in decision-making process.

Copyright © 2005, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.